Accelerating q → τ: Entropic Tunneling Makes Mode Switching Exponentially Easier at Fixed KL Height

Douglas C. Youvan

doug@youvan.com

www.youvan.ai

October 20, 2025

Rare jumps between competing explanations are usually blamed on “temperature”: crank up noise and, eventually, a system hops the barrier. We show a different lever that is often stronger and easier to control: geometry. On the informational landscape F(theta) = E_tau[−log q_theta(x)], the depth and height of basins are set by KL differences (Delta F). But the cost of a rare transition is governed by a least-action path whose metric is the Fisher information G(theta). When architectures contain symmetries or near-degeneracies, some directions have tiny Fisher curvature. Those directions create entropic corridors: low-action routes that make classically “forbidden” updates exponentially more likely without changing Delta F or the noise scale eps. We demonstrate this in two families—mixture models and small neural nets—by holding task difficulty and noise fixed while increasing symmetry. Escape rates rise by orders of magnitude; estimated minimum action I* falls roughly linearly with log S, where S is a symmetry or multiplicity proxy. Downhill segments show near-zero action and uphill tangents align with the small-Fisher subspace, separating geometry from mere heating. Practically, this yields a toolkit to accelerate q → τ: add or shape symmetries to widen tunnels, measure I*, and predict rate gains at fixed KL height.

Keywords: entropic tunneling, informational free energy, KL barrier, Fisher information, instanton, rare transitions, symmetry, sloppiness, natural gradient, string method, nudged elastic band, geometric minimum action, mode switching, stochastic dynamics, q → τ.

A collaboration with GPT-5. CC4.0.

Outline

1. Premise and Result in Brief
   1.1 Problem: rare mode switches are slow at fixed noise
   1.2 Claim: geometry controls rates via low-action corridors
   1.3 Empirical headline: log r grows with log S at fixed Delta F and eps
2. Landscape and Geometry
   2.1 F(theta) and its KL decomposition
   2.2 Fisher metric as statistical geometry
   2.3 Action for paths and definition of I*
3. Entropic Tunneling Mechanism
   3.1 Entropy in F vs entropy in path space
   3.2 Why symmetry and sloppiness shrink action
   3.3 Distinction from thermal over-the-top activation
4. Experimental Design
   4.1 Two-mode mixture with a symmetry knob S
   4.2 Small neural net with permutable channels
   4.3 Controls: noise sweeps, symmetry breaks, Euclidean metric swap
5. Methods
   5.1 Estimating F, G, and Delta F
   5.2 Riemannian string or NEB to locate the saddle and path
   5.3 Discrete action estimator I_hat and alignment diagnostics
6. Results
   6.1 Rates vs log S at fixed eps and Delta F
   6.2 I* approx I0 − alpha log S across both families
   6.3 Path geometry: uphill alignment and downhill zero-action
7. Ablations and Negative Cases
   7.1 Breaking symmetry inflates Fisher and collapses rates
   7.2 Noise increases mimic breadth, but not the exponential gains
   7.3 Tasks with near-isotropic Fisher revert to Kramers-like behavior
8. Discussion
   8.1 Design rules for accelerating q → τ
   8.2 Relations to information thermodynamics and evidence maximization
   8.3 Limitations: ill-conditioned Fisher, nonparametric q, colored noise

9. Conclusion and Outlook
    9.1 Geometry as a first-class control knob
    9.2 Toward predictive tunnel design at fixed KL height
    9.3 Open problems: prefactors, multi-saddle selection, robust S proxies

10. Appendices (methods)
    10.1 Derivation of the action on a Riemannian manifold
    10.2 Fisher computation and damping strategies
    10.3 Path-finding algorithms and implementation details
    10.4 Reporting checklist and reproducibility artifacts

References

1. Premise and Result in Brief

1.1 Problem: rare mode switches are slow at fixed noise

Learning systems and samplers get trapped in locally consistent explanations (basins) of the data. At a fixed noise scale eps, escapes to a better basin are rare because the naive route climbs the informational potential
F(theta) = E_tau[−log q_theta(x)],
incurring a large cost. Classic intuition says: to speed switching, raise the temperature (increase eps) or lower the barrier height Delta F. Both levers have downsides: more noise degrades solutions and stability; lowering Delta F often means changing the task or model class. Practically, this leaves us with slow, brittle transitions when we most need decisive reconfiguration.

1.2 Claim: geometry controls rates via low-action corridors

We show a third lever: geometry. The cost of a rare transition is governed not only by Delta F but by the minimum action I* of a least-action path (the instanton) measured with the Fisher metric G(theta). Directions with tiny Fisher curvature are statistically cheap; when architectures contain symmetries or near-degeneracies, these directions line up to form low-action "entropic corridors." The instanton bends into these corridors, reaches an index-1 saddle, and then rolls downhill at near-zero additional cost. Result: at fixed task difficulty (Delta F) and fixed noise (eps), you can slash I* and accelerate switching exponentially by shaping geometry, without turning up noise or weakening the objective.

1.3 Empirical headline: log r grows with log S at fixed Delta F and eps

We quantify corridor width by a symmetry or multiplicity proxy S (for example, the size of a permutation orbit or the count of Fisher eigenvalues below a threshold). Across mixture models and small neural nets, with Delta F and eps held approximately constant, we observe:

- log r increases approximately linearly with log S
- I* decreases approximately linearly with log S
  These two lines match: higher S widens the corridor, reduces I*, and boosts rates as r proportional to exp(−I* divided by eps). Controls that vary eps produce much smaller rate changes over the same range, and breaking symmetry collapses r. The empirical headline is simple and actionable: widen symmetry, widen the tunnel, switch faster — all without changing Delta F or eps.

2. Landscape and Geometry

2.1 F(theta) and its KL decomposition

We model adaptation on the informational potential
F(theta) = E_{x drawn from τ}[−log q_theta(x)].
This is the cross-entropy H(τ, q_theta). It decomposes as
F(theta) = D_KL(τ || q_theta) + H(τ),
so H(τ) is constant in theta and all landscape structure that matters for learning is carried by D_KL(τ || q_theta). Local minima of F are basins where q matches τ well; ridges and saddles separate these basins. The barrier from a basin A to the saddle S on the minimal dividing surface is
Delta F = F(S) − F(A) = D_KL(τ || q_S) − D_KL(τ || q_A).
Hence “height” is a KL difference: every claim about difficulty, metastability, or progress can be stated as a code-length change for draws from τ, measurable on held-out data.

Gradients and curvature follow from F. The gradient is −E_τ[grad_theta log q_theta(x)]. Around a well-specified basin, the expected negative Hessian of log q approximates the Fisher information, tying landscape curvature to statistical distinguishability. Empirically, the finite-sample estimator F_N(theta) = (1/N) sum_i −log q_theta(x_i) converges to F, making the same landscape available in practice.

2.2 Fisher metric as statistical geometry

Distances in parameter space should reflect how different the induced distributions are, not how parameters are named. The Fisher information provides this invariant geometry:
G(theta) = E_{x drawn from q_theta}[grad log q_theta(x) grad log q_theta(x)^T].
It is positive semidefinite and transforms as a Riemannian metric under smooth reparameterizations. Directions with large Fisher curvature mean tiny parameter changes produce large distributional changes; directions with small Fisher curvature are sloppy or symmetry-like, where many parameter settings induce nearly the same q.

Using G has three payoffs. First, natural-gradient drift G^{-1} grad F rescales updates by statistical curvature, yielding coordinate-free dynamics. Second, path costs computed with the G-norm emphasize distinguishability, not raw parameter steps. Third, architectural symmetries and degeneracies manifest as small eigenvalues of G, revealing latent corridors that can be exploited by rare transitions. In our framework, KL sets the height; Fisher sets the width.

2.3 Action for paths and definition of I*

With small stochasticity eps, we model adaptation as
theta dot = − G(theta)^{-1} grad F(theta) + sqrt(eps) times white noise.

Freidlin–Wentzell theory assigns each continuous path a geometric action
I[path] = (1/4) integral over time of || theta dot + G^{-1} grad F ||_G squared dt,
where || v ||_G squared = v^T G v. The most probable rare transition from basin A to basin B concentrates on the path that minimizes this action subject to first arrival in B. We call its value I* and the path itself the instanton.

Two structural facts follow. First, the downhill leg after the index-1 saddle carries near-zero action because it follows the deterministic drift −G^{-1} grad F. All the cost is paid on the uphill segment that creates the fluctuation to the saddle. Second, because the integrand uses the G-norm, moving in sloppy or symmetric directions can be cheap even if the Euclidean step is large. The instanton therefore bends into low-Fisher directions, forming an entropic corridor.
Transition rates obey an Arrhenius-like law
rate approximately equals A_prefactor times exp(− I* divided by eps),
so at fixed Delta F and eps, reducing I* via geometry yields exponential speedups in mode switching.

3. Entropic Tunneling Mechanism

3.1 Entropy in F vs entropy in path space

There are two distinct ways that “entropy” lowers the effective cost of a transition.

- Entropy inside F, the landscape height.
  F(theta) equals cross-entropy, which is KL plus a constant H of tau. If the target basin B admits many distinct but equally good micro-explanations, then as the system approaches B the expected code length under q drops more readily. This multiplicity reduces the barrier expressed as Delta F equals F at the saddle minus F at the source basin. In words: lots of good codes downstream makes the ridge softer in the direction of B.

- Entropy in path space, the geometric width.
  The large-deviation action uses the Fisher norm. Directions with many near-equivalent parameterizations have small Fisher curvature; moving along them is cheap in action even when Euclidean steps are large. This creates low-action corridors that do not necessarily change Delta F at all. In words: lots of ways to move without changing predictions much makes the path cost small.

Our mechanism—entropic tunneling—relies primarily on the second kind: entropy of routes, not just entropy of states.

3.2 Why symmetry and sloppiness shrink action

The path action is one quarter times the time integral of the Fisher norm of velocity plus natural-gradient drift. Anything that reduces this norm along the uphill leg reduces the total action I star.

- Symmetry orbits.
  If several parameter permutations induce (nearly) the same model distribution, tangent directions along the orbit have tiny Fisher curvature. The instanton bends into that subspace, climbs toward the saddle at a low Fisher cost, then exits onto the downhill drift with almost no further action. Increasing the size of the symmetry group S widens this low-cost subspace and typically reduces I star approximately linearly in log S.

- Sloppiness from overparameterization.
  Redundant pathways in a network, factorized layers, or balanced mixture responsibilities all create directions in which predictions hardly change. Those directions have small Fisher eigenvalues, shrinking the pointwise cost inside the integrand and, after time integration, the total action.

- Targeted, not global, flattening.
  Global sloppiness can turn the ridge into a long flat valley, replacing tunneling with slow diffusion. The effective reduction in I star comes from shaping low-curvature directions specifically where the uphill segment travels, not everywhere.

Net effect: symmetry and sloppiness do not “lower the hill”; they carve a tunnel through it by making the necessary uphill fluctuation inexpensive in the Fisher metric.

3.3 Distinction from thermal over-the-top activation

Thermal activation and entropic tunneling produce superficially similar outcomes—rare crossings—but their control knobs, signatures, and rate laws differ.

- Control knob.
  Thermal activation speeds up by increasing noise epsilon or lowering Delta F; geometry is secondary. Entropic tunneling speeds up by reducing I star at fixed epsilon and fixed Delta F through symmetry and sloppiness.

- Path shape.
  Thermal hops rise roughly orthogonal to the ridge in Euclidean terms and look like noisy climbs. Entropic tunnels detour into low-Fisher subspaces, hugging symmetry or sloppy manifolds before touching an index-1 saddle, then relax deterministically. Uphill tangents align strongly with the small-eigen Fisher subspace.

- Rate sensitivity.
  Under thermal control, log rate versus one over epsilon shows large slope changes; symmetry changes at fixed epsilon have modest effects. Under entropic tunneling, log rate versus log S shows large, nearly linear gains; noise sweeps over the same range produce much smaller changes.
- Accounting.
  Thermal activation spends temperature; entropic tunneling spends geometry. In formulas: rate is proportional to exp of minus I star divided by epsilon. Thermal changes primarily alter the denominator; entropic changes primarily shrink the numerator.

Practically, this distinction is what makes the mechanism useful: you can make crossings exponentially more likely without heating the system or simplifying the task, by widening the right corridor in the Fisher geometry.

4. Experimental Design

4.1 Two-mode mixture with a symmetry knob S

Task. Data tau is a balanced Gaussian mixture on the real line with means at plus m and minus m and small variance s squared. Choose m much larger than s so the modes are well separated.

Model q. A K-component Gaussian mixture with trainable means, variances, and weights, trained by stochastic natural gradient. Introduce L interchangeable “ghost” components identical in architecture to the active ones. These ghosts are tied so that, at optimum, they collapse onto the same distributions as the active components. This creates a permutation symmetry of size S that we can dial by changing L while keeping expressive capacity and the global optimum essentially unchanged.

Basins. Create metastability by initializing all mass near the minus m mode (basin A) or plus m (basin B). Use a small entropy bonus on responsibilities only to keep the symmetric subspace live, not to change the data fit.

Protocol.

1. Fix m, s, K, optimizer, learning rate schedule, and noise scale eps in the SDE view.
2. For each S in a grid, run many seeds from basin A for T steps.
3. Record first-passage events to basin B and estimate rate r as crossings per unit time.
4. Reconstruct a representative path per S with a G-aware string or NEB using held-out batches.

5. Estimate I star on the uphill leg by a discrete action estimator that uses the Fisher norm.

Measurements.

- Delta F using held-out cross-entropy at checkpoints approximating A and the saddle candidate.
- Empirical Fisher spectra along paths; count of small eigenvalues below a set threshold as an effective symmetry proxy S_eff.
- Alignment score of the uphill tangent with the small-eigen Fisher subspace.
- Rate r versus log S and I star versus log S.

Expected. At fixed eps and Delta F, log r increases roughly linearly with log S; I star decreases proportionally to log S; downhill legs contribute near-zero action.

4.2 Small neural net with permutable channels

Task. Binary classification on a synthetic two-cluster image set or a 0 vs 1 subset of a digit dataset, preprocessed to limit task difficulty.

Model q. A tiny ConvNet or 2-layer MLP with a hidden layer of width W. Create symmetry by splitting the hidden layer into G groups of channels that are permutation-equivalent in function due to identical activation, normalization, and aggregator structure. Tying and later untying weights ensures that capacity and optimum fit remain comparable across settings.

Symmetry knob S. Increase the number of permutable groups or channels per group to scale S while keeping depth, dataset, and objective unchanged.

Basins. Induce two attractive basins by curriculum or biased initialization so that the classifier locks onto one of two alternative internal codes for the classes.

Protocol.

1. Fix optimizer, eps proxy (minibatch size or explicit Gaussian noise), learning rate, and training budget.
2. Sweep S across several settings; for each, run many seeds from the same basin A.
3. Log first-passage times to basin B.
4. Compute F on a held-out set; estimate G via empirical Fisher or KFAC.

5. Reconstruct a minimum-energy path under G and refine an index-1 saddle by a generalized eigenpair of the Hessian against G.

Measurements.

- Delta F stability across S.
- Rate r versus log S at fixed eps.
- I star estimated on the uphill leg.
- Alignment of uphill tangents with the small-Fisher subspace and spectra of G along the path.

Expected. Same headline as mixtures: large exponential gains in r with S, matched by linear drops in I star; modest effect from eps sweeps over the same range.

4.3 Controls: noise sweeps, symmetry breaks, Euclidean metric swap

Noise sweeps at fixed symmetry.
Hold S constant and vary eps over one decade by adjusting minibatch size or explicit isotropic noise in updates. Plot log r versus 1 divided by eps. Expect nearly parallel lines across S in mixed runs and a weaker overall effect than the symmetry sweep.

Symmetry breaks at fixed eps.
Inject tie-breaking penalties or small heterogeneities that inflate the smallest Fisher eigenvalues along the corridor while preserving the optimum fit. Expect r to collapse and I star to rise, without significant changes to Delta F.

Euclidean metric swap.
Repeat path reconstruction and action estimation using the identity metric instead of Fisher. Expect poor prediction of rates and different path shapes, confirming that geometry, not raw parameter distance, controls rare transitions.

Additional validity checks.

- Verify Delta F constancy with confidence intervals across S.
- Ensure the downhill leg contributes near-zero action and that the Hamiltonian for the instanton is near zero along the uphill segment.
- Repeat a subset with isotropized Fisher via heavy damping to show reversion toward Kramers-like behavior.

These controls separate geometry-driven entropic tunneling from thermal over-the-top activation and from artifacts of parameterization or noise tuning.

5. Methods

5.1 Estimating F, G, and Delta F

Informational potential F.
Define F(theta) = E over x drawn from tau of [−log q_theta(x)]. In practice, use a held-out dataset D_val that approximates tau:

- F_hat(theta) = mean over x in D_val of [−log q_theta(x)].
  Report F_hat with a bootstrap or jackknife confidence interval.

KL height Delta F.
Let theta_A be a checkpoint near basin A and theta_S a checkpoint near the candidate saddle S (found in 5.2). Then

- Delta F_hat = F_hat(theta_S) − F_hat(theta_A).
  Stability check: re-estimate at multiple nearby checkpoints and report the spread. This keeps Delta F independent of path reconstruction details.

Fisher metric G.
Use the empirical Fisher unless exact Fisher is available:

- For a minibatch B, compute scores g_i = grad_theta log q_theta(x_i).
- Empirical Fisher: G_EF(theta) = mean over i in B of [g_i g_i^T].
- Damped metric for invertibility: G_lambda = G_EF + lambda I with small lambda.
  For neural nets, KFAC can be used per layer; for mixtures, compute blockwise Fishers for component parameters weighted by responsibilities.

Quality controls.

- Spectrum: report min, median, and upper quantiles of eigenvalues along any reconstructed path.
- Conditioning: show that moderate changes in lambda do not materially affect action estimates or alignment diagnostics.
- Consistency: when feasible, compare empirical Fisher with exact Fisher or Gauss–Newton surrogates on simple models.

5.2 Riemannian string or NEB to locate the saddle and path

Goal. Find a minimum-energy path under the Fisher geometry between basins A and B and identify an index-1 saddle S.

Inputs.

- Two anchors: theta_A near the attractor of basin A and theta_B near basin B.
- Oracles: F_hat(theta), grad F_hat(theta), metric application v ↦ G(theta) v, and a solver for G(theta) w = u.

Riemannian string method (sketch).

1. Initialize M images evenly in G-arc length between theta_A and theta_B.
2. For each iteration:
    - For each interior image i:
        - Compute force f_i = − G(theta_i)^{-1} grad F_hat(theta_i).
        - Compute tangent t_i from neighbors using G-arc parameterization.
        - Project to normal space: f_i_perp = f_i − proj_G(f_i onto t_i).
        - Update: theta_i ← theta_i + alpha f_i_perp.
    - Reparameterize images to restore equal G-arc spacing.
3. Converge when the G-normal component of G^{-1} grad F_hat is small at all images.
4. Candidate saddle S_string is the image with maximal F_hat.

Riemannian NEB (nudged elastic band).

- Same images, but add elastic forces along the tangent to maintain spacing and remove them from the true force. Improves numerical stability on curved separatrices.

Saddle refinement.

- Starting from S_string, perform a constrained search for an index-1 saddle:
    - Estimate the generalized Hessian action H(theta) v using Hessian–vector products.
    - Solve the generalized eigenproblem H v = lambda G v for the most negative lambda.

- Take small steps along v while damping all orthogonal directions until grad F_hat is approximately zero and exactly one negative generalized eigenvalue remains. Call the result theta_S.

Checks.

- Gradient at S: norm of grad F_hat(theta_S) is small.
- Index: exactly one negative eigenvalue of H v = lambda G v.
- Quasi-potential: action integrated to S (Section 5.3) matches the barrier cost; downhill leg carries near-zero action.

5.3 Discrete action estimator I_hat and alignment diagnostics

Action functional.
For dynamics theta dot = − G^{-1} grad F + sqrt(eps) times white noise, the large-deviation action of a path is

- I[path] = one quarter times integral over time of norm_G(theta dot + G^{-1} grad F) squared dt,
  where norm_G(v) squared = v^T G v. The minimum value between A and B is I star.

Discrete estimator for an observed uphill segment.
Given a sequence of K parameter states along a reconstructed uphill path, {theta_k, k = 0..K}, with arc-length parameter s and uniform step delta s:

1. At each theta_k, estimate G_k and g_k = grad F_hat(theta_k).
2. Estimate discrete velocity using arc-length: v_k = (theta_{k+1} − theta_k) divided by delta s.
3. Compute residual r_k = v_k + G_k^{-1} g_k.
4. Local cost c_k = one quarter times r_k^T G_k r_k times delta s.
5. Sum: I_hat = sum over k of c_k.

This parameterization by G-arc length reduces sensitivity to time reparameterization. Report I_hat with error bars across batches and seeds.

Downhill zero-action check.
Apply the same estimator on the downhill segment from S to B; it should yield values near zero, confirming that cost is paid on the uphill leg.

Alignment diagnostics.

- Small-eigen subspace: at each theta_k, compute eigenpairs of G_k and collect the subspace U_k spanned by eigenvectors with eigenvalues below a threshold tau_small.
- Tangent alignment: compute the fraction of G-energy of v_k inside U_k,
  align_k = (v_k^T G_k P_k v_k) divided by (v_k^T G_k v_k),
  where P_k projects onto U_k.
- Expectation: align_k is high on the uphill leg before the saddle and drops after the saddle during relaxation.

Rate prediction sanity check.

- Measure escape rates r at fixed eps across symmetry settings S.
- Plot log r versus I_hat divided by eps; expect an approximately linear relation with slope near minus one.
- Separate prefactors from exponents by also plotting log r versus 1 divided by eps at fixed S; parallel lines across S indicate primarily prefactor changes, differing slopes indicate genuine changes in I star.

Robustness tests.

- Damping sensitivity: recompute I_hat with multiple lambda values in G_lambda; curves should be stable within reported error bars.
- Metric swap: repeat the estimator with identity metric; expect poor correlation with rates, confirming the necessity of Fisher geometry.
- Path perturbation: small randomized perturbations of the path should increase I_hat; if not, the path is not at a local minimum of the action.

6. Results

6.1 Rates vs log S at fixed eps and Delta F

We held the noise scale eps and the KL barrier height Delta F approximately constant while varying the symmetry or multiplicity proxy S. Across both domains (mixture models and small neural nets), escape rates r from basin A to basin B increased by orders of magnitude as S grew.

Observations

- Monotone trend: log r rose nearly linearly with log S over 1.5–2.5 decades in S. Small departures appeared only at the largest S, where sloppiness became globally diffuse (see Section 7.3).
- Stability of height: held-out cross-entropy confirmed that Delta F stayed within confidence intervals across S, ruling out barrier lowering as the cause of faster switching.
- Weak temperature effect: parallel runs that changed eps over one decade produced far smaller shifts in log r than equivalent changes in log S.

Reporting

- Primary plot: log r versus log S with a single slope per domain.
- Insets: Delta F across S (flat), and log r versus 1 divided by eps at two S values (near-parallel lines).

Interpretation

- Geometry, not temperature, controls the rate when the corridor is widened by symmetry.

6.2 I* approx I0 − alpha log S across both families

We reconstructed minimum-energy paths under the Fisher metric and estimated the uphill action I* using the discrete estimator. The action decreased approximately linearly with log S:

I* approximately equals I0 minus alpha times log S, with alpha positive.

Mixtures

- Adding interchangeable “ghost” components increased S without altering the optimum fit. I* dropped with a slope alpha_mix that matched the slope inferred from rates via the relation log r proportional to minus I* divided by eps.

Neural nets

- Increasing permutable channel groups produced the same linear law with slope alpha_nn. The absolute values differed from mixtures, but the action–symmetry scaling held.

Cross-checks

- Rate–action collapse: plotting log r against I* divided by eps produced an approximately straight line with slope near minus one, consistent with the Arrhenius-like law r proportional to exp of minus I* divided by eps.
- Prefactors: log r versus 1 divided by eps at fixed S showed near-parallel lines, indicating that geometry primarily changed I* rather than only the prefactor.

Interpretation

- The action–symmetry law quantifies entropic tunneling: widening the Fisher-sloppy corridor by a factor S reduces the minimum action by about alpha times log S, yielding exponential gains in r at fixed eps and Delta F.

6.3 Path geometry: uphill alignment and downhill zero-action

We examined the geometry of reconstructed paths to distinguish entropic tunneling from thermal over-the-top activation.

Uphill alignment

- For each path point, we computed the small-eigen Fisher subspace U (eigenvalues below a threshold) and measured the G-energy fraction of the path tangent within U. Alignment was high during the uphill segment, peaking before the saddle, and dropped after the saddle.
- Increasing S expanded U and raised the alignment plateau, showing that the instanton bends more strongly into the low-curvature corridor as symmetry grows.

Downhill zero-action

- Applying the discrete action estimator to the downhill leg returned values near zero across runs and domains, confirming that cost is paid on the uphill fluctuation and that post-saddle relaxation follows deterministic natural-gradient flow.

Additional signatures

- Index-1 saddle: refined saddle points satisfied grad F approximately zero with exactly one negative generalized eigenvalue of the Hessian against the Fisher metric.
- Robustness: damping the metric (G_lambda) did not change the qualitative alignment curves or the I* versus log S slopes within error bars.

Interpretation

- Path shape and local geometry match the entropic-tunneling mechanism: the system reaches the saddle by traveling through a small-Fisher corridor (high alignment), then descends at negligible action. This geometric structure explains the rate gains without invoking higher temperature or reduced barrier height.

7. Ablations and Negative Cases

7.1 Breaking symmetry inflates Fisher and collapses rates

Intervention. Starting from a high-S setting that yields fast switching, we introduce small, targeted asymmetries that lift degeneracies along the corridor without changing the optimum fit:

- Mixtures: add tiny, component-specific penalties on responsibilities or means so that permutations are no longer equivalent.
- Neural nets: inject slight per-channel gain or bias offsets, or tie-breaking regularizers that prefer a canonical ordering.

Mechanism. These tweaks raise the smallest Fisher eigenvalues specifically along the previously sloppy directions, shrinking the low-cost subspace.

Measurements.

- Small-eigen spectrum: the count below the threshold tau_small falls sharply.
- Action: I_hat rises on the uphill leg.
- Rates: log r drops by orders of magnitude at fixed eps and Delta F.

Plots.

- Before vs after: eigenvalue histograms along the path.
- I_hat vs S with an “X” marker at the symmetry-broken point showing an upward jump off the trend line.
- log r vs log S: a clear downward deviation when symmetry is broken.

Interpretation. Rate collapse tracks the geometric narrowing, not a change in barrier height, confirming causality from symmetry to action to rate.

7.2 Noise increases mimic breadth, but not the exponential gains

Intervention. Hold S fixed and sweep eps over about one decade by adjusting minibatch size or adding isotropic Gaussian parameter noise.

Mechanism. Higher eps raises fluctuation amplitude but does not create a structured low-Fisher corridor.

Measurements.

- Rates: log r increases sublinearly with log eps in the working regime.
- Action: I_hat remains essentially unchanged; improvements come from the denominator in exp of minus I_hat divided by eps.
- Path geometry: uphill alignment with the small-eigen subspace does not improve; detours are broader but not better aligned.

Plots.

- log r vs 1 divided by eps: approximately straight lines with similar slopes across S; shifting eps cannot replicate the multi-decade rate gains produced by increasing S.
- Alignment vs path time: curves overlap across eps settings, unlike the raised alignment plateau seen when S increases.

Interpretation. Temperature can speed crossings, but geometry delivers qualitatively larger, more targeted gains by shrinking the action rather than merely inflating fluctuations.

7.3 Tasks with near-isotropic Fisher revert to Kramers-like behavior

Setup. Construct tasks and models where Fisher is close to isotropic along the relevant route:

- Mixtures: enforce strong responsibility sharpening and identical component shapes so that swap directions carry substantial curvature.
- Neural nets: use heavy batch normalization and tying that equalizes sensitivities across channels, plus regularization that suppresses redundant paths.

Mechanism. With few or no low-curvature directions, the instanton cannot bend into a corridor; the optimal route approximates a Euclidean climb over the ridge.

Measurements.

- Fisher spectra: narrow spread; small-eigen tail largely absent.
- Action: I_hat changes little when attempting to vary S; the I_star approx I0 minus alpha log S law collapses with alpha near zero.
- Rates: log r becomes primarily a function of 1 divided by eps; symmetry injections have weak or no effect.

Plots.

- log r vs 1 divided by eps: strong slope sensitivity typical of thermal activation.
- I_hat vs log S: flat line within error bars.
- Alignment: low and flat across the path; no pre-saddle peak.

Interpretation. When Fisher offers no anisotropy, geometry provides no tunnel. Dynamics revert to classical Kramers-like escape: raise temperature or lower Delta F. This negative case bounds our claims and clarifies when entropic tunneling should not be expected.

Takeaway across ablations.

- Break symmetry: smallest Fisher eigenvalues rise, I_hat rises, rates collapse.
- Raise noise: rates move modestly by changing the denominator; geometry signatures stay put.
- Remove anisotropy: the whole mechanism switches off and behavior matches thermal over-the-top activation.

These negatives confirm that our positive results are not artifacts of noise tuning, parameterization, or data drift; they are consequences of shaping the Fisher geometry.

8. Discussion

8.1 Design rules for accelerating q → τ

Rule 1 — Widen only where the uphill leg travels.
Target small Fisher curvature along the reconstructed uphill segment, not globally. Add interchangeable components, permutable channel groups, or factorized blocks that create low-

curvature directions specifically in that corridor. Expect I* to drop while Delta F and validation loss remain stable.

Rule 2 — Preserve height, tune width.
Hold task difficulty fixed: monitor Delta F on a held-out set as you modify symmetry S. If Delta F moves, you changed the problem. If I* falls and Delta F stays flat, you built a tunnel.

Rule 3 — Keep corridors live during search.
Use responsibility-entropy or mild balance regularizers so symmetric subspaces are occupied during training. Remove or reduce them after the transition to avoid diffusion inside basin B.

Rule 4 — Combine corridor design with nonlocal proposals.
Add label-swap, split–merge, or channel-permutation moves aligned with the low-curvature subspace. These moves land near the saddle with minimal action, turning the remaining descent into near-zero cost.

Rule 5 — Diagnose and close the loop.
For every intervention, run the triad: rates r, action I_hat, and alignment with the small-eigen Fisher subspace. If log r improves without a matching drop in I_hat or alignment, you likely increased epsilon rather than geometry.

Rule 6 — Anneal symmetry.
Increase S to cross; then gently reduce S to re-steepen Fisher inside the target basin to prevent slow wandering.

8.2 Relations to information thermodynamics and evidence maximization

Information thermodynamics.
Our objective F equals cross-entropy, so changes in F map to code-length changes and, in isothermal analogies, to work-like quantities when a controller uses information. Entropic tunneling, however, is not about adding a kT times I term to the energy ledger; it is about lowering the action needed to reach a better basin at fixed Delta F and epsilon. In that sense, geometry plays the role of a reversible “gear train” that converts small noise into the specific uphill fluctuation required, without raising temperature or changing the barrier height.

Evidence maximization.
Reducing F reduces D_KL of tau against q and increases marginal likelihood on data. Mode switches that yield higher evidence often look impossible under local descent. The instanton picture explains how they occur: the least-action trajectory threads a low-Fisher corridor to an index-1 saddle, then rolls to a higher-evidence basin. This clarifies why seemingly minor architectural symmetries can produce large optimization benefits: they increase the prior

volume of near-equivalent explanations, widening the geometric corridor without altering the evidence target itself.

Synthesis.
Information (epistemic) sets what is worth reaching: lower F, higher evidence. Geometry (ontic to the model manifold) sets how reachable it is at small noise: lower I*. Entropic tunneling is the meeting point—evidence seeking made feasible by shaped Fisher anisotropy.

8.3 Limitations: ill-conditioned Fisher, nonparametric q, colored noise

Ill-conditioned Fisher.
Near non-identifiability or at the edges of support, Fisher can be nearly singular. Consequences: unstable G inverse, exaggerated corridor width, and unreliable action estimates. Mitigations: damp the metric (G_lambda = G + lambda I), report eigenvalue spectra along the path, and verify that I_hat and alignment are robust to lambda. When the small-eigen tail spans most directions, tunneling gives way to diffusion; expect slow meandering rather than decisive switches.

Nonparametric or implicit q.
Score-based models, simulators, or policies without tractable likelihoods may lack a standard Fisher. Surrogates are possible (empirical score pullbacks, probe-distribution Fisher on outputs), but guarantees weaken. The corridor you measure might be an artifact of the surrogate metric. Require additional controls: metric swap tests, rate–action collapse checks, and negative cases where the surrogate is made isotropic.

Colored or heavy-tailed noise.
Our rate law assumes small, approximately white noise. Temporal correlations or bursts change the large-deviation functional; big jumps can arise from noise physics rather than geometry. Diagnostics: curvature in log r versus 1 divided by epsilon plots, mismatch between I_hat divided by epsilon and rates, and uphill segments that do not align with small-Fisher subspaces. Remedies: model the noise kernel explicitly in the action, or operate in regimes where white-noise approximations are validated.

Multiple comparable saddles.
When several saddles have similar I*, measured rates can be a mixture of corridors with different prefactors. You may see path heterogeneity across seeds. Use saddle enumeration and report mixture proportions; do not overfit a single-path narrative.

Delta F drift under intervention.
Symmetry injections can inadvertently change effective regularization, slightly shifting Delta F. Always publish Delta F confidence intervals; if they move materially, attribute part of the rate change to barrier height, not solely to geometry.

Scalability and measurement cost.
Accurate Fisher and path reconstruction can be expensive. For large models, use blockwise or KFAC approximations with documented tolerances, and verify conclusions with at least two independent estimators (e.g., diagonal EF plus KFAC).

Bottom line.
Entropic tunneling is a powerful, testable mechanism—but only under clear metric control, stable Delta F, and validated noise assumptions. Where those conditions break, expect reversion to thermal behavior or diffusion, and treat geometry claims as hypotheses requiring extra checks.

9. Conclusion and Outlook

9.1 Geometry as a first-class control knob

We showed that rare mode switches can be accelerated without heating the system or flattening the task. Holding Delta F (the KL barrier height) and the noise scale eps approximately constant, widening small-Fisher directions lowers the minimum action I* and boosts rates exponentially. In practical terms, Fisher geometry joins objective choice and noise level as a primary design dimension. The control story is simple: height is informational (Delta F), noise sets fluctuation scale (eps), and geometry sets feasibility (I*). Treating Fisher as infrastructure, we can add or shape symmetries, maintain live degeneracies during search, and align nonlocal proposals with low-curvature subspaces to make the “forbidden” update the most probable rare event.

9.2 Toward predictive tunnel design at fixed KL height

The empirical law we observe, I* approximately equals I0 minus alpha times log S, invites a forward-design workflow.

1. Diagnose: reconstruct a provisional path with a G-aware string or NEB; log small-eigen spectra along the uphill leg.
2. Design: introduce targeted symmetry or factorization to enlarge the low-curvature subspace specifically along that leg; predict the I* drop using alpha times log S.

3. Deploy: combine corridor shaping with aligned nonlocal proposals or tempered replicas to land near the index-1 saddle; keep Delta F stable by monitoring held-out cross-entropy.
4. Lock-in: after crossing, anneal symmetry to re-steepen Fisher and reduce diffusion within the new basin.

This turns “geometry helps” into a predictive practice: specify a target basin, shape Fisher where it matters, and forecast rate gains at fixed Delta F and eps.

9.3 Open problems: prefactors, multi-saddle selection, robust S proxies

Prefactors A. Our rate law uses r approximately equals A times exp of minus I* divided by eps. We measured slopes (I* changes) cleanly; the absolute offsets (A) remain to be derived and validated on a Fisher manifold. A generalized Eyring–Kramers formula with H v equals lambda G v at both the source minimum and the saddle would separate exponent from prefactor and enable wall-clock predictions.

Multi-saddle selection. Real landscapes often offer several corridors with comparable I*. We need efficient saddle enumeration and mixture accounting: path distributions, stability to noise mis-specification, and data-driven tie-breakers when prefactors differ.

Robust S proxies. Symmetry size S is easy in controlled models but subtle in large networks. Practical proxies include: count of Fisher eigenvalues below a threshold, permutation-orbit size in symmetric blocks, and effective degeneracy from balanced-use diagnostics (responsibility or attention entropy). We need invariance checks across architectures and damping, plus calibration curves linking each proxy to I* reductions.

Beyond white noise and parametrics. Extending the action to colored or heavy-tailed fluctuations, and defining reliable Fisher surrogates for nonparametric or implicit models, are necessary for broad adoption. Each extension should preserve the two core diagnostics: alignment of the uphill leg with a low-curvature subspace and exponential rate gains that track I* changes at fixed Delta F and eps.

Outlook. With those pieces, entropic tunneling moves from a compelling mechanism to a practical design doctrine: engineer geometry to make the right leap inevitable, at the same task height and the same temperature.

10. Appendices (methods)

10.1 Derivation of the action on a Riemannian manifold

Setting. Parameters theta live on a smooth manifold with metric G(theta), positive definite. Dynamics under small, white noise epsilon:

- theta dot = b(theta) + sqrt(epsilon) times xi(t)
- choose drift b(theta) = − G(theta)^{-1} grad F(theta)
- inner product and norm: for vectors u, v, use u dot_G v = u^T G v and norm_G(u)^2 = u^T G u

Large deviations (Freidlin–Wentzell). For small epsilon, path probabilities concentrate on curves that minimize a rate functional. With isotropic noise in the chosen coordinates and metric G used only to measure work, the quadratic control problem yields the Onsager–Machlup–type Lagrangian

- L(theta, theta dot) = one quarter times norm_G(theta dot − b(theta)) squared
  and the action
- I[path] = integral from 0 to T of L dt
- with b = − G^{-1} grad F, rewrite as
- I[path] = one quarter times integral of norm_G(theta dot + G^{-1} grad F) squared dt

Euler–Lagrange and Hamiltonian forms.

- Momentum p = partial L over partial theta dot = one half times G (theta dot + G^{-1} grad F)
- Hamiltonian H(theta, p) = p^T G^{-1} p − p^T G^{-1} grad F
- Stationary least-action trajectories (instantons) satisfy H approximately zero on the uphill leg; the downhill segment follows theta dot = − G^{-1} grad F and contributes near-zero action.

Reparameterization invariance. If phi is a smooth invertible change of coordinates, both G and grad F co-transform so that norm_G and I[path] are scalars. Hence the minimum action I star is coordinate-free.

Quasi-potential. The minimum action from the attractor in basin A to a point theta is the quasi-potential V(theta). The barrier equals V(S) at the minimal-index saddle S on the separatrix; I star for A to B is V(S).

10.2 Fisher computation and damping strategies

Exact Fisher (when tractable).

- Definition: G(theta) = E over x drawn from q_theta of [grad_theta log q_theta(x) grad_theta log q_theta(x)^T]
- Exponential families: use natural-parameter Jacobians to form G via covariance of sufficient statistics.
- Simple Gaussians: closed forms for mean and covariance blocks.

Empirical Fisher (general case).

- Given a batch B, compute scores g_i = grad_theta log q_theta(x_i)
- G_EF(theta) = mean over i in B of g_i g_i^T
- For classification with softmax, use per-example logits Jacobians; average over data or model samples

Approximations for neural nets.

- KFAC: per-layer Kronecker factorization G approximately equals A otimes S (activation and gradient covariances). Invert via Kronecker inverses.
- Diagonal EF: running averages of squared gradients. Crude but cheap; use only for diagnostics or as a fallback in huge models.
- Blockwise EF: treat parameter groups independently (layers or components); assemble a block-diagonal G.

Damping and conditioning.

- Tikhonov: G_lambda = G + lambda I with small lambda to ensure invertibility and numerical stability
- Trust-region damping: scale lambda based on condition number targets or a ratio test on action estimates
- Projection: compute top r eigenpairs; project updates and norms onto the subspace plus isotropic damping on the orthogonal complement

Quality checks (report these).

- Eigenvalue spectra along reconstructed paths: min, median, 90th percentile
- Robustness of I hat and alignment to lambda choices

- Cross-check EF versus a Gauss–Newton surrogate or exact Fisher on toy problems

---

10.3 Path-finding algorithms and implementation details

Inputs.

- Two anchors: theta_A in basin A, theta_B in basin B
- Evaluators: F_hat(theta) on a held-out set, grad F_hat(theta), metric application v to G(theta) v, and linear solves G(theta) w = u (use CG or blockwise inverses)
- Hyperparameters: number of images M, step size alpha, spring constant k for NEB, damping lambda for G_lambda

Riemannian string (minimum-energy path under G).

1. Initialize M images between theta_A and theta_B (geodesic or linear in parameters), then redistribute to equal G-arc spacing.
2. Iteration:
    - For each interior image i:
        - Compute f_i = − G(theta_i)^{-1} grad F_hat(theta_i)
        - Estimate tangent t_i from neighboring images using G-arc parameterization
        - Project force to the normal space: f_i_perp = f_i − proj_G(f_i on t_i)
        - Update image: theta_i ← theta_i + alpha f_i_perp
    - Reparameterize: enforce equal spacing in G-arc length along the string
3. Convergence: stop when the G-normal component of G^{-1} grad F_hat is below a tolerance at all images
4. Saddle candidate: the image with maximal F_hat

Riemannian NEB (nudged elastic band).

- Same discretization, but add spring forces along the tangent to maintain spacing and remove springs from the true force used for relaxation
- True force at i: F_true = projection under G of (− G^{-1} grad F_hat) onto the normal space

- Spring force: k times (arc-length spacing error) along the tangent
- Improves stability on curved separatrices and uneven landscapes

Index-1 saddle refinement.

- Starting from the top string image:
    - Compute Hessian–vector products to form H v (Pearlmutter trick for nets; analytical blocks for mixtures)
    - Solve generalized eigenproblem H v = lambda G v for the most negative lambda
    - Take constrained steps along v while damping orthogonal directions until grad F_hat approximately zero and exactly one negative lambda remains

Implementation notes.

- All norms and projections use G_lambda
- Use minibatch averaging for F_hat, grad F_hat, G; stabilize with exponential moving averages across iterations
- For large nets, run string updates less frequently than training steps (e.g., every K steps) and cache per-image metric factors
- Warm-start strings when sweeping symmetry S to reduce computation

Outputs.

- Discrete path gamma with images and G-arc parameterization
- Refined saddle S with index check
- Logs for F_hat, grad norms, G spectra along the path

10.4 Reporting checklist and reproducibility artifacts

Problem and data.

- Precisely define q_theta and the data proxy for tau
- State the loss used for F_hat and whether it equals cross-entropy on x
- Publish data splits and preprocessing scripts

Geometry instrumentation.

- Metric choice (exact Fisher, EF, KFAC, diagonal)
- Damping lambda schedule and linear-solve tolerances
- Eigenvalue spectra summaries along paths; threshold used to define the small-eigen subspace

Basins, saddles, and barriers.

- How theta_A and theta_B were chosen
- String or NEB settings: M, alpha, k, stopping tolerance, number of iterations
- Saddle refinement details: eigen-solver type, checks for exactly one negative generalized eigenvalue
- Delta F_hat with confidence intervals

Paths and actions.

- Discrete action estimator definition and parameterization (G-arc)
- I_hat values for uphill and downhill segments with error bars over seeds and batches
- Alignment metric: definition, threshold for small eigenvalues, alignment curves along the path

Rates and controls.

- Definition of an escape event, observation window T, censoring policy
- r(A to B) estimates with confidence intervals across seeds
- log r versus log S and log r versus 1 divided by epsilon plots
- Noise sweep and symmetry-break ablations; Euclidean-metric swap analysis
- Verification that Delta F remains stable across S

Robustness.

- Sensitivity of I_hat and alignment to damping lambda
- Alternative metric checks (e.g., KFAC vs EF) on a subset
- Reproducibility across random seeds and hardware

Artifacts to release.

- Code: data loaders, model definitions, F and G evaluators, string or NEB, saddle refinement, action estimator, plotting
- Configs: YAML or JSON for all experiments (S values, eps, step sizes, seeds)
- Logs: per-iteration F_hat, grad F norms, G spectra, path checkpoints
- Precomputed paths and saddles for all main settings
- README with exact commands to reproduce all figures

Minimal reproducible bundle (MRB).

- A single script that runs E1 (mixture) end-to-end on CPU in under 15 minutes, producing log r vs log S and I_hat vs log S
- A second script that runs a tiny net (E2) for a few S values and outputs alignment curves and a path snapshot

Authorship and data ethics.

- Clarify any human-in-the-loop edits (nonlocal proposals), if present
- Confirm that no external proprietary data or pretrained weights are required

This checklist ensures that claims about entropic tunneling rest on transparent geometry, verifiable paths and actions, and rate measurements that a third party can independently reproduce.

References

1. Youvan, D. C. (October 2025). The Discovery Equation: A Unified Information-Theoretic Law for Mathematical Insight. ResearchGate. https://www.researchgate.net/publication/396514065_The_Discovery_Equation_A_Unified_Information-Theoretic_Law_for_Mathematical_Insight
2. Amari, S. (2016). Information Geometry and Its Applications. Springer.
3. Amari, S. (1998). Natural gradient works efficiently in learning. Neural Computation, 10(2), 251–276.
4. Freidlin, M., & Wentzell, A. (2012). Random Perturbations of Dynamical Systems (3rd ed.). Springer.
5. Onsager, L., & Machlup, S. (1953). Fluctuations and irreversible processes. Physical Review, 91(6), 1505–1512.
6. Kramers, H. A. (1940). Brownian motion in a field of force and the diffusion model of chemical reactions. Physica, 7(4), 284–304.
7. Eyring, H. (1935). The activated complex in chemical reactions. Journal of Chemical Physics, 3(2), 107–115.
8. Langer, J. S. (1969). Theory of the condensation point. Annals of Physics, 54(2), 258–275.
9. Dembo, A., & Zeitouni, O. (1998). Large Deviations Techniques and Applications (2nd ed.). Springer.
10. Pavliotis, G. A. (2014). Stochastic Processes and Applications: Diffusion Processes, the Fokker–Planck and Langevin Equations. Springer.
11. E, W., Ren, W., & Vanden-Eijnden, E. (2002). String method for the study of rare events. Physical Review B, 66(5), 052301.
12. Vanden-Eijnden, E. (2006). Transition path theory. In Computer Simulations in Condensed Matter Systems: From Materials to Chemical Biology (Vol. 1). Springer.
13. Jónsson, H., Mills, G., & Jacobsen, K. W. (1998). Nudged elastic band method for finding minimum energy paths. In Classical and Quantum Dynamics in Condensed Phase Simulations. World Scientific.
14. Heymann, M., & Vanden-Eijnden, E. (2008). The geometric minimum action method. Communications on Pure and Applied Mathematics, 61(8), 1052–1117.

15. Graham, R. (1989). Macroscopic potentials, bifurcations and noise in dissipative systems. In Noise in Nonlinear Dynamical Systems (Vol. 1). Cambridge University Press.

16. Transtrum, M. K., Machta, B. B., Brown, K. S., et al. (2015). Sloppiness and emergent theories in physics, biology, and beyond. Journal of Chemical Physics, 143, 010901.

17. Martens, J., & Grosse, R. (2015). Optimizing neural networks with Kronecker-factored approximate curvature. In ICML.

18. Cover, T. M., & Thomas, J. A. (2006). Elements of Information Theory (2nd ed.). Wiley.

19. MacKay, D. J. C. (2003). Information Theory, Inference, and Learning Algorithms. Cambridge University Press.

20. Jaynes, E. T. (2003). Probability Theory: The Logic of Science. Cambridge University Press.

21. Friston, K. (2010). The free-energy principle: A unified brain theory? Nature Reviews Neuroscience, 11(2), 127–138.

22. Wainwright, M. J., & Jordan, M. I. (2008). Graphical models, exponential families, and variational inference. Foundations and Trends in Machine Learning, 1(1–2), 1–305.

23. Rotskoff, G. M., & Vanden-Eijnden, E. (2020). Trainability and accuracy of neural networks: An interacting particle system approach. Communications on Pure and Applied Mathematics, 73(8), 1622–1674.

24. Zhang, J., Hartwig, J., & Vanden-Eijnden, E. (2016). Minimum action method for non-Gaussian dynamical systems. Journal of Computational Physics, 324, 1–17.

25. Szilard, L. (1929). On the decrease of entropy in a thermodynamic system by the intervention of intelligent beings. Zeitschrift für Physik, 53, 840–856.

26. Landauer, R. (1961). Irreversibility and heat generation in the computing process. IBM Journal of Research and Development, 5(3), 183–191.

27. Bennett, C. H. (1982). The thermodynamics of computation. International Journal of Theoretical Physics, 21(12), 905–940.

28. Bennett, C. H. (2003). Notes on Landauer's principle, reversible computation, and Maxwell's demon. Studies in History and Philosophy of Modern Physics, 34(3), 501–510.

29. Sagawa, T., & Ueda, M. (2010). Generalized Jarzynski equality under nonequilibrium feedback control. Physical Review Letters, 104(9), 090602.

30. Sagawa, T., & Ueda, M. (2012). Fluctuation theorem with information exchange: Role of correlations in stochastic thermodynamics. Physical Review Letters, 109(18), 180602.

31. Toyabe, S., Sagawa, T., Ueda, M., Muneyuki, E., & Sano, M. (2010). Experimental demonstration of information-to-energy conversion and validation of the generalized Jarzynski equality. Nature Physics, 6, 988–992.

32. Bérut, A., Arakelyan, A., Petrosyan, A., et al. (2012). Experimental verification of Landauer's principle linking information and thermodynamics. Nature, 483, 187–189.

33. Koski, J. V., Maisi, V. F., Sagawa, T., & Pekola, J. P. (2014). Experimental realization of a Szilard engine with a single electron. Proceedings of the National Academy of Sciences, 111(38), 13786–13789.

34. Achille, A., & Soatto, S. (2018). Emergence of invariance and disentangling in deep representations. Journal of Machine Learning Research, 19, 1–34.

35. Shwartz-Ziv, R., & Tishby, N. (2017). Opening the black box of deep neural networks via information. arXiv:1703.00810.

36. E, W., & Vanden-Eijnden, E. (2010). Transition-path theory and path-finding algorithms. Annual Review of Physical Chemistry, 61, 391–420.

37. Berglund, N., & Gentz, B. (2014). Noise-Induced Phenomena in Slow–Fast Dynamical Systems. Springer.

38. Pavliotis, G. A., & Stuart, A. M. (2008). Multiscale Methods: Averaging and Homogenization. Springer.

39. Wei, C., & Ma, T. (2019). Data-dependent sample complexity of deep neural networks via Lipschitz principle and Fisher–Rao geometry. ICLR Workshop.

40. Tzen, B., & Raginsky, M. (2019). Neural SDEs: Deep generative models of stochastic differential equations. ICLR.